The Development of Geospatially-enabled Grid Technology for Earth Science

Liping Di
Laboratory for Advanced Information Technology and Standards (LAITS)
George Mason University
9801 Greenbelt Road, Suite 316-317
Lanham, MD 20706, USA
ldi@gmu.edu

AIST-02-0160

Integration of Grid and OGC Technologies for Earth Science Modeling and Applications

PI, Liping Di, LAITS/GMU.
Co-I, Williams Johnston NASA Ames and DOE LBNL.
Co-I, Deans Williams, DOE LLNL.

Introduction

- Geospatial data is the major type of data that human beings has collected.
  - more than 80% of the data are geospatial data.
- Image/gridded data is dominant form of geospatial data in terms of volume.
  - Most of those data are collected by the EO community.
- Geospatial data will grow to ~exabyte very soon.
  - NASA EOSDIS has more than three petabyte of data in archives; more than 2 terabytes per day of new data are added.
  - Application data centers: 10's of terabytes of imagery
  - Tens of thousands of datasets on-line now.
- How to effectively, wisely, and easily use the geospatial data is the key information technology issue that we have to solve.

The Problems

- In order for the geospatial data to be useful, they have to be converted to user-specific information and knowledge.
- However, the conversion requires:
  - Significant amount of knowledge
    - Domain knowledge for information/knowledge extraction from raw data
    - Domain knowledge on the geospatial data processing/formats
  - Significant amount of computer hardware and software resources.
  - As a result, currently the use of geospatial data is very expensive
- Most geo-imagery will never be directly analysed by humans
  - Human attention is the scarce resource, insufficient to analyse petabytes of geospatial data.
  - Many datasets have not been analysed once before they are archived.
- The fundamental problem is that current data and information systems running by EO agencies only can provide data at best, not the user-specific information and knowledge.
  - Rich in geospatial data but poor in up-to-date geospatial information and knowledge.

What the User Needs

- The ready-to-use geospatial information and knowledge that can answer the specific application questions of the end users.
  - It is not important whether an answer is derived from Landsat TM, SPOT, or field observations, as long as the users can easily obtain the right answer at adequate accuracy from geospatial information systems.
  - Not only experts can use the geospatial data, everyone, from students to decision-makers can obtain and use the geospatial information and knowledge easily.
- A system that can automatically convert the geospatial data to user-specific geoinformation and knowledge.
  - Automate the process from geospatial data to information to knowledge.

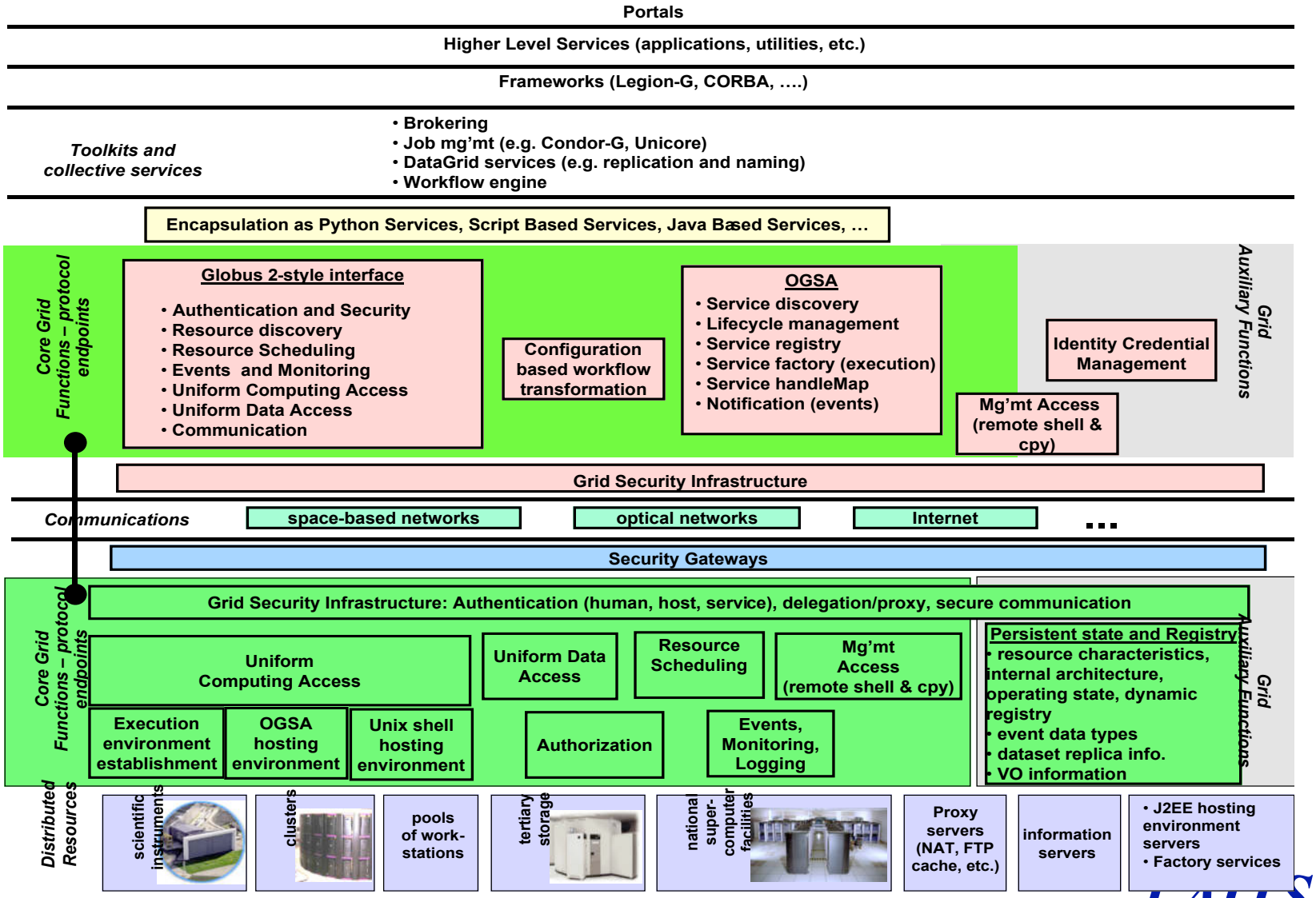
The Grid Technology

- The Grid technology is developed for secured computational resource sharing and coordinated problem solving in dynamic, multi-institutional virtual organizations.
  - Computer CPU cycles
  - Storage
  - Networks.
  - Data, Information, algorithms, software, services.
  - Human expertise.
- It was originally motivated and supported from sciences and engineering requiring high-end computing, for sharing geographically distributed high-end computing resources.
- The core of the technology is the the open source middleware called Globus Toolkit.
  - The latest version of Globus is version 3.2 which implements the Open Grid Service Architecture (OGSA)

What Grid Provides

- Enabling new large-scale scientific research and applications through the coordinated use of geographically distributed resources
  - E.g., distributed collaboration, data access and analysis, distributed computing
- Persistent infrastructure for Grid computing
  - E.g., certificate authorities and policies, protocols for resource discovery/access

The Layered Grid Architecture



Why Grid is useful to the EO community?

- Earth observation community is one of the key communities for collecting, managing, processing, archiving and distribution geospatial data and information.
- Because of the large volumes of EO data and geographically scattered receiving and processing facilities, the EO data and associated computational resources are naturally distributed.
- The multi-discipline nature of global change research and remote sensing applications requires the integrated analysis of huge volume of multi-source data from multiple data centers. This requires sharing of both data and computing powers among data centers.
- Therefore, Grid is an ideal technology for EO community.

Why geospatial extensions of Grid is needed

- Geospatial data and information are significantly different from those in other disciplines.
  - Very complex and diverse.
    - Formats, projection, resolutions.
    - Hyper-dimensions: spatial, temporal, spectral, thematic.
    - Raster vs. vectors
  - Large data volume
    - more than 80% of data human beings has collected is spatial data.
- The geospatial community has developed a set of standards specifically for geospatial data and information that users have been familiar with. (e.g., OGC, ISO, FGDC).
- Grid technology is developed for general sharing of computational resources and not aware of the specialty of geospatial data.
- In order to make Grid technology applicable to geospatial data, we have to do the geospatial domain-specific extensions.

Areas of Extensions

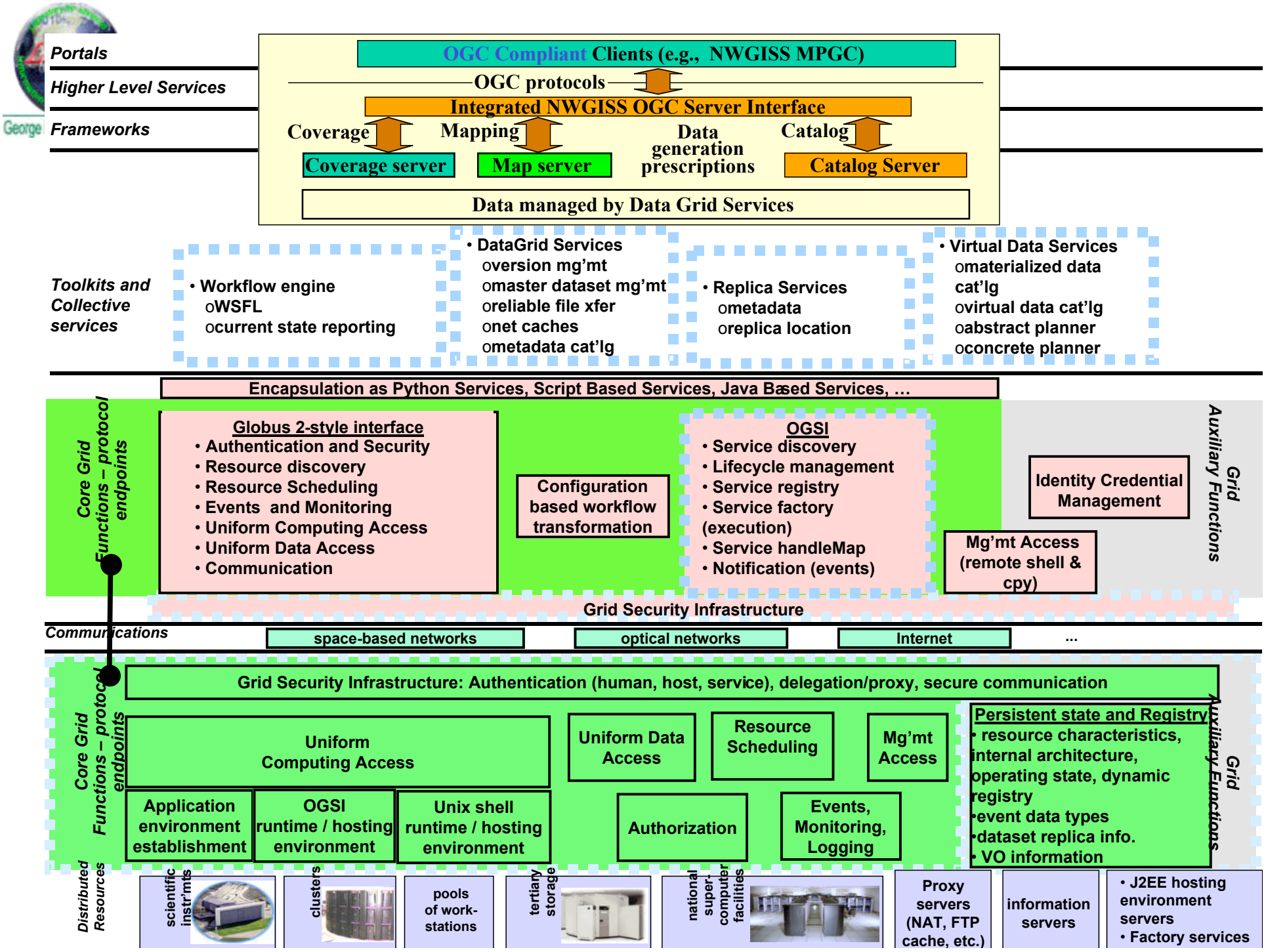
- Internally in the Grid, it have to be spatially aware.
  - Extend Globus toolkit to handle the spatial, spectral, temporal, thematic based spatial data and information management.
  - Develop enough Grid-enable tools for geospatial data handling/services.
- Must provide data/information access and services interfaces that are standard in the geospatial community.
  - The Open GIS Consortium's Web Data Access/Service interfaces (e.g., OGC WCS, WMS, WFS, and WRS).

The OGC Web Service Specifications

- The Web Coverage Services (WCS) specification: defines the standard interfaces between web-based clients and servers for accessing coverage data.
  - All imagery type of remote sensing data is coverage data.
- The Web Feature Services (WFS) specification: defines the standard interfaces between web-based clients and servers for accessing feature-based geospatial data.
  - vector and point data are feature data.
- The Web Map Services (WMS) specification: define the standard interfaces for accessing and assembling maps from multiple servers.
  - visualization of geospatial data
- The Web Registries Services (WRS) specification: defines the interfaces between web-based clients and servers for finding the required data or services from registries.
- WCS, WFS, WRS, and WMS form the foundation for the interoperable geospatial data access and service environment

Project Objectives

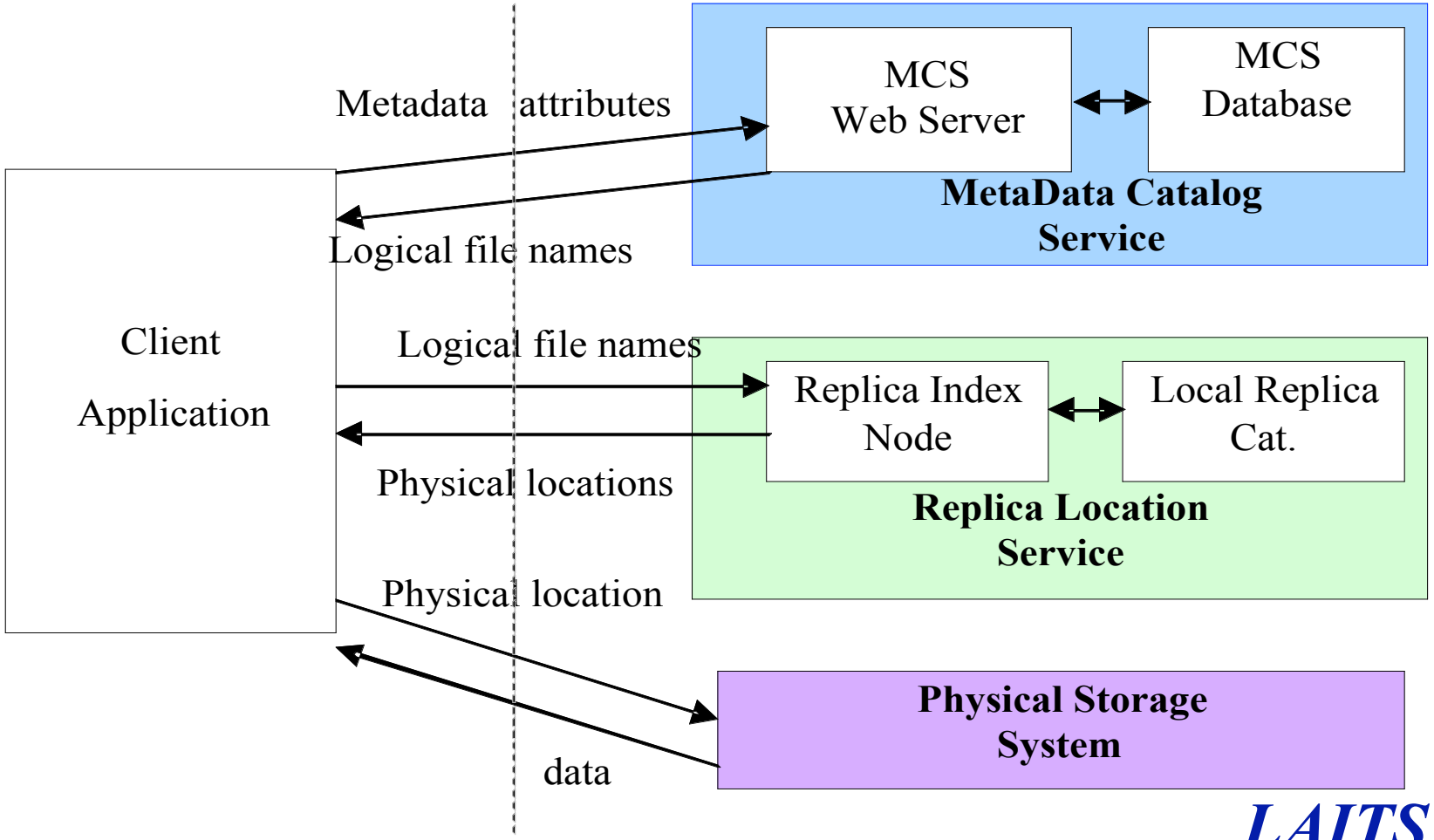
- Making NASA EOSDIS data easily accessible to Earth science modeling and applications communities by integrating the advantages of both OGC and Grid technologies.
- Making Grid technology geospatially enabled and OGC standard compliant and making OGC technology Grid enabled.
- Allowing researchers to focus on science and not issues with computing, storage and bandwidth resources, and data receipt, data format and data set manipulation,
- Achieving the access to NASA EOSDIS Data Pools and ESG (Earth System Grid).
- Overall, development of Geospatial Grid technology
  - Geospatial extensions of Grid tech.


Portals
Higher Level Services
Frameworks
OGC Compliant Clients (e.g., NWGISS MPGC)
OGC protocols
Integrated NWGISS OGC Server Interface
Coverage
Mapping
Data generation prescriptions
Catalog
Coverage server
Map server
Catalog Server
Data managed by Data Grid Services
Toolkits and Collective services
• Workflow engine
oWSFL
ocurrent state reporting
• DataGrid Services
oversion mg'mt
omaster dataset mg'mt
oreliable file xfer
onet caches
ometadata cat'lg
• Replica Services
ometadata
oreplica location
• Virtual Data Services
omaterialized data cat'lg
ovirtual data cat'lg
oabstract planner
oconcrete planner
Encapsulation as Python Services, Script Based Services, Java Based Services, …
Core Grid Functions – protocol endpoints
Globus 2-style interface
• Authentication and Security
• Resource discovery
• Resource Scheduling
• Events and Monitoring
• Uniform Computing Access
• Uniform Data Access
• Communication
Configuration based workflow transformation
OGSI
• Service discovery
• Lifecycle management
• Service registry
• Service factory (execution)
• Service handleMap
• Notification (events)
Identity Credential Management
Mg'mt Access (remote shell & cpy)
Grid Auxiliary Functions
Grid Security Infrastructure
Communications
space-based networks
optical networks
Internet
...
Grid Security Infrastructure: Authentication (human, host, service), delegation/proxy, secure communication
Uniform Computing Access
Uniform Data Access
Resource Scheduling
Mg'mt Access
Application environment establishment
OGSI runtime / hosting environment
Unix shell runtime / hosting environment
Authorization
Events, Monitoring, Logging
Persistent state and Registry
• resource characteristics, internal architecture, operating state, dynamic registry
•event data types
•dataset replica info.
• VO information
Distributed Resources
scientific instr'mts
clusters
pools of work-stations
tertiary storage
national super-computer facilities
Proxy servers (NAT, FTP cache, etc.)
information servers
• J2EE hosting environment servers
• Factory services

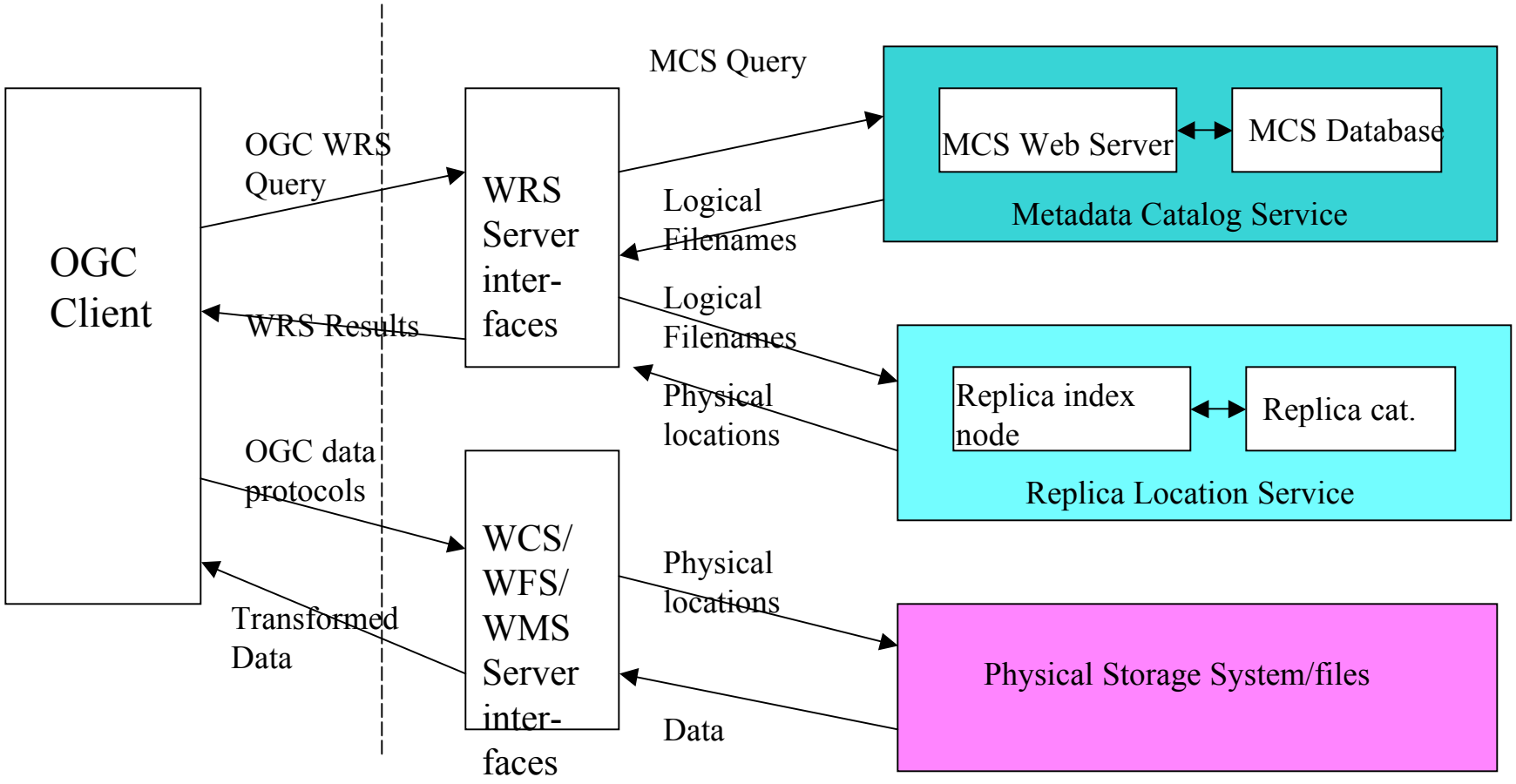
Implementation Plan

- The *first phase* is the testbed and initial integration, including the setup of the development environment, preliminary design of the integration, and implementation of WCS access to Grid-managed data.
- The *second phase* is the data naming and location transparency, which include the use of Data Grid and Replica Services (metadata catalogues, replication location management, reliable file transfer services, and network caches) to provide naming and location independence for data used by NWGISS and revising NWGISS to invoke such Grid services.
  - The approach to investigating the Data Grid and Replica Services will be to configure a Data Grid testbed. This will be followed by the integration of NWGISS data catalogs into a data Grid catalog and the investigation of naming approaches, followed by interfacing NWGISS with data generators and Data Grid Replica Location service
- The *third phase* is the virtual dataset research and development.

Data Access Sequences in the Data Grid



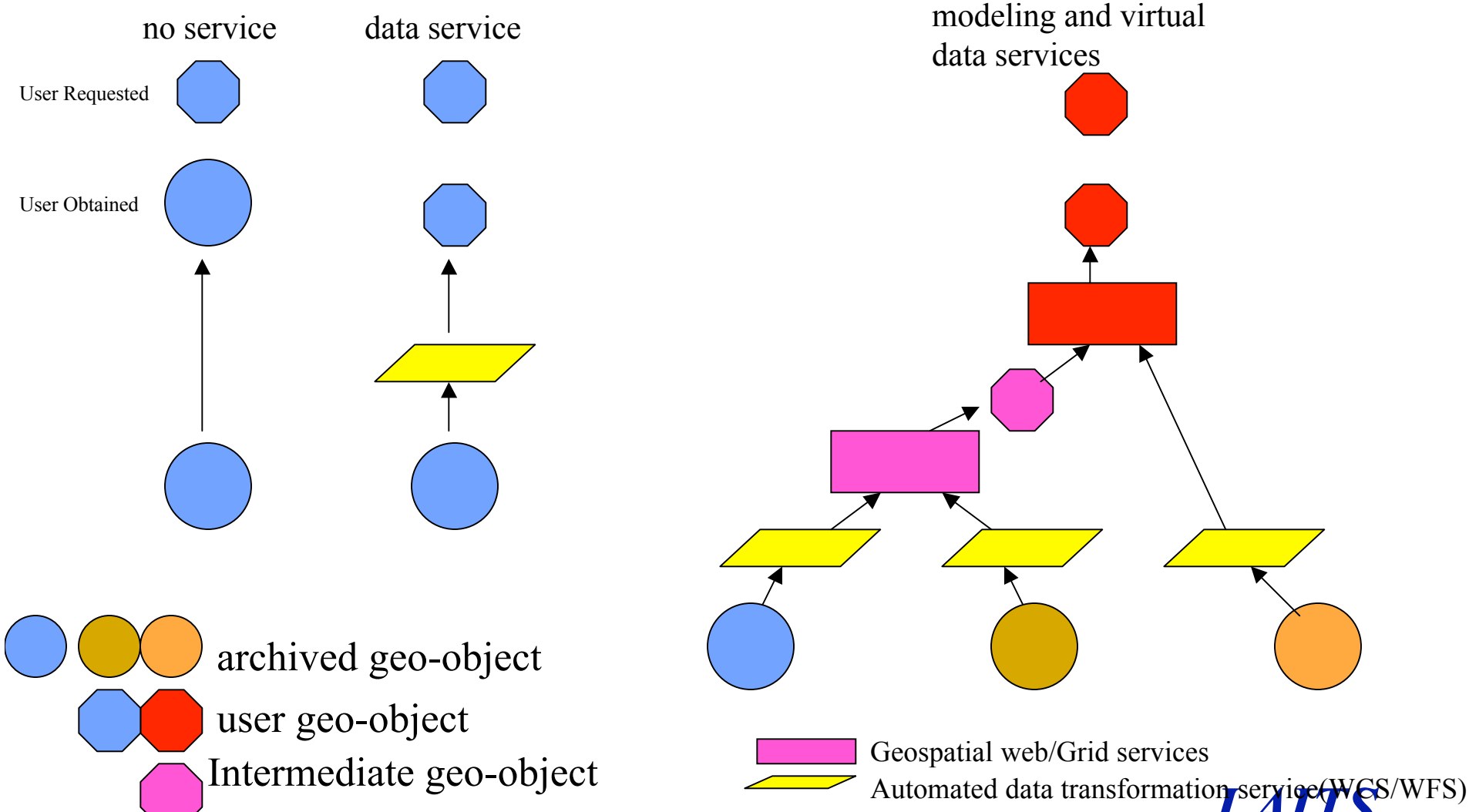
OGC Interfaces to the Geospatial Data Grid



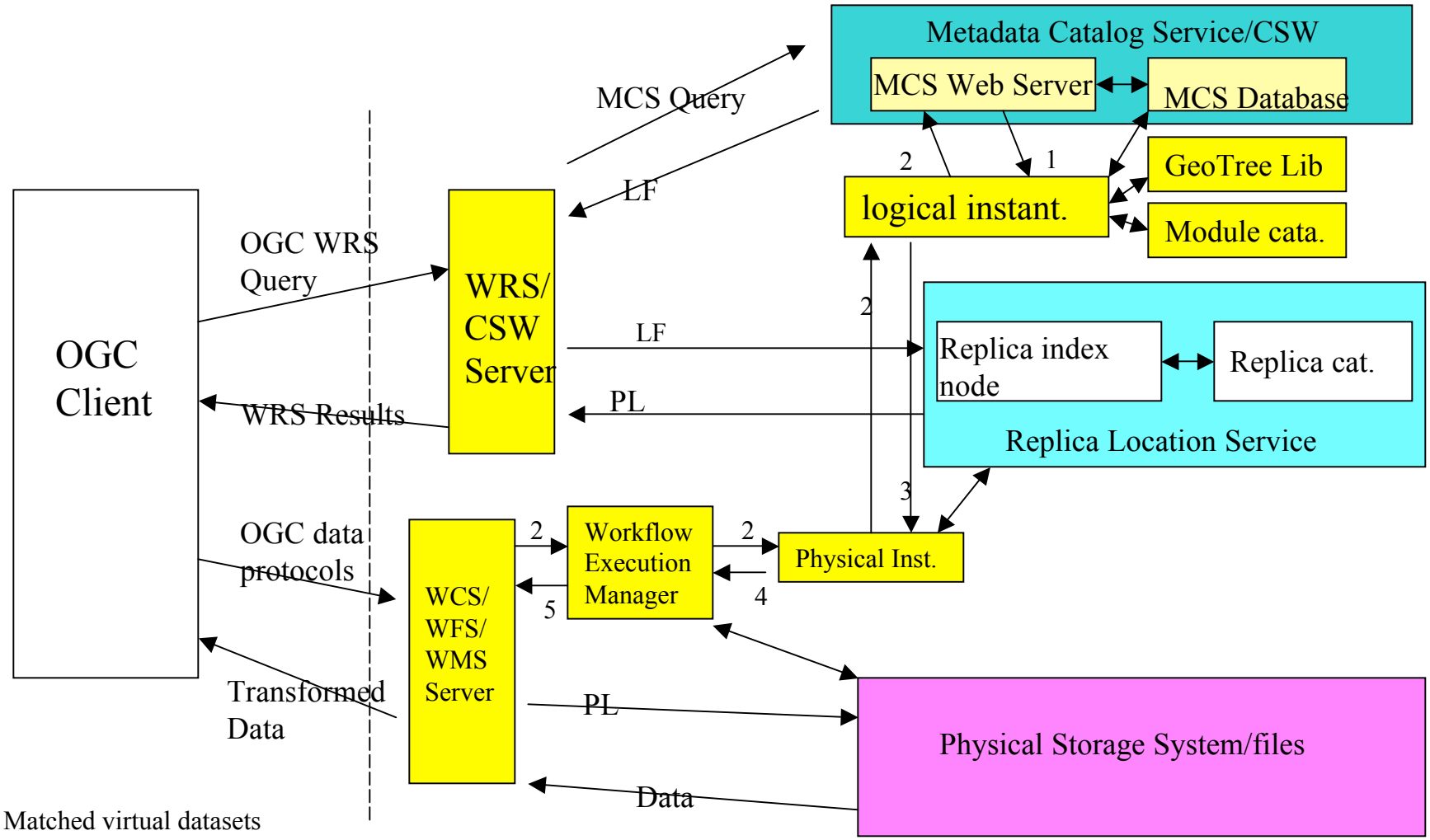
Virtual datasets

- A virtual dataset is a dataset that:
  - not exist in a data and information system
  - The system knows how to create it on-demand.
  - A virtual dataset, once created, can be kept for fulfilling the same request from next users.
- The client/data user will not know the difference between a real dataset and a virtual dataset.
- A virtual dataset can be produced (materialized) by
  - running a program dedicated to the production of the virtual dataset (dedicated program approach).
  - running a series of service modules, each one takes care of a small step of the materialization of the virtual dataset (service approach).

Geo-object, Geo-tree, Virtual Dataset, Geospatial Models



Virtual Data Services In the Geospatial Data Grid



1: Matched virtual datasets
2: logically instanced virtual filenames (LIVF)
3. logical workflow
4. Physical workflow
5. Data

LF: Logical filename
PL: Physical Location

Yellow: New component
Light Yellow: Modified MCS component

Current Status

- Established the testbed environments:
  - Installed Grid software (Globus Toolkits) at GMU LAITS and NASA Ames.
  - Established the GMU LAITS VO and NASA IPG VO for testbed
  - Provided multiple version of OGC compliant software (NWGISS) for various kinds of Operating System on computers of testbed.
- Designed the system architecture of the integration of OGC with Grid Technology.
- Designed and implemented the Spatial Grid Service (SGS) as an agent between OGC compliant software and Grid software.
- Implemented the NWGISS WCS access to Grid-managed data based on Grid's MCS/RLS and MySQL database.

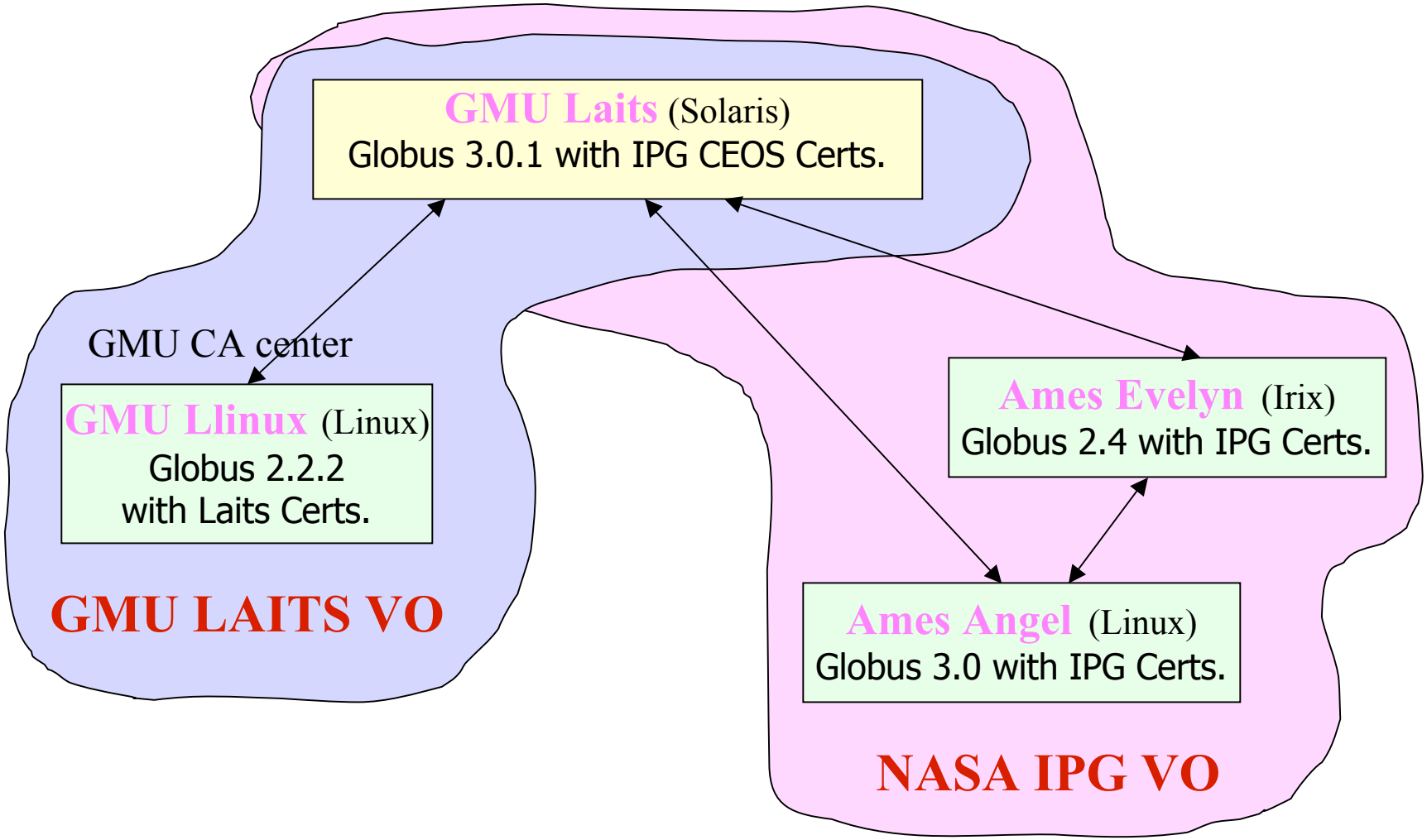
Current Status –Cont'd

- Designed and implemented the OGC Web Registry Service (WRS) based on the Grid technology (Globus 2.2) for the goal of OGC technology Grid enabled. (http://llinux.laits.gmu.edu:8080/WRS)
- Utilized the Grid Security Infrastructure (GSI) (CA's certificates for VO) to make NWGISS running based on a secure Grid framework.
- Built a massive storage device (2TB) and populate with NASA EOSDIS data.
- Designed and implemented the OGC Catalog Service for Web (CSW) as a Web Service, which will be further advanced to work as a Catalog Service for geospatial Grids conforming to Grid technology and specifications. (http://www.laits.gmu.edu:8099/csw)

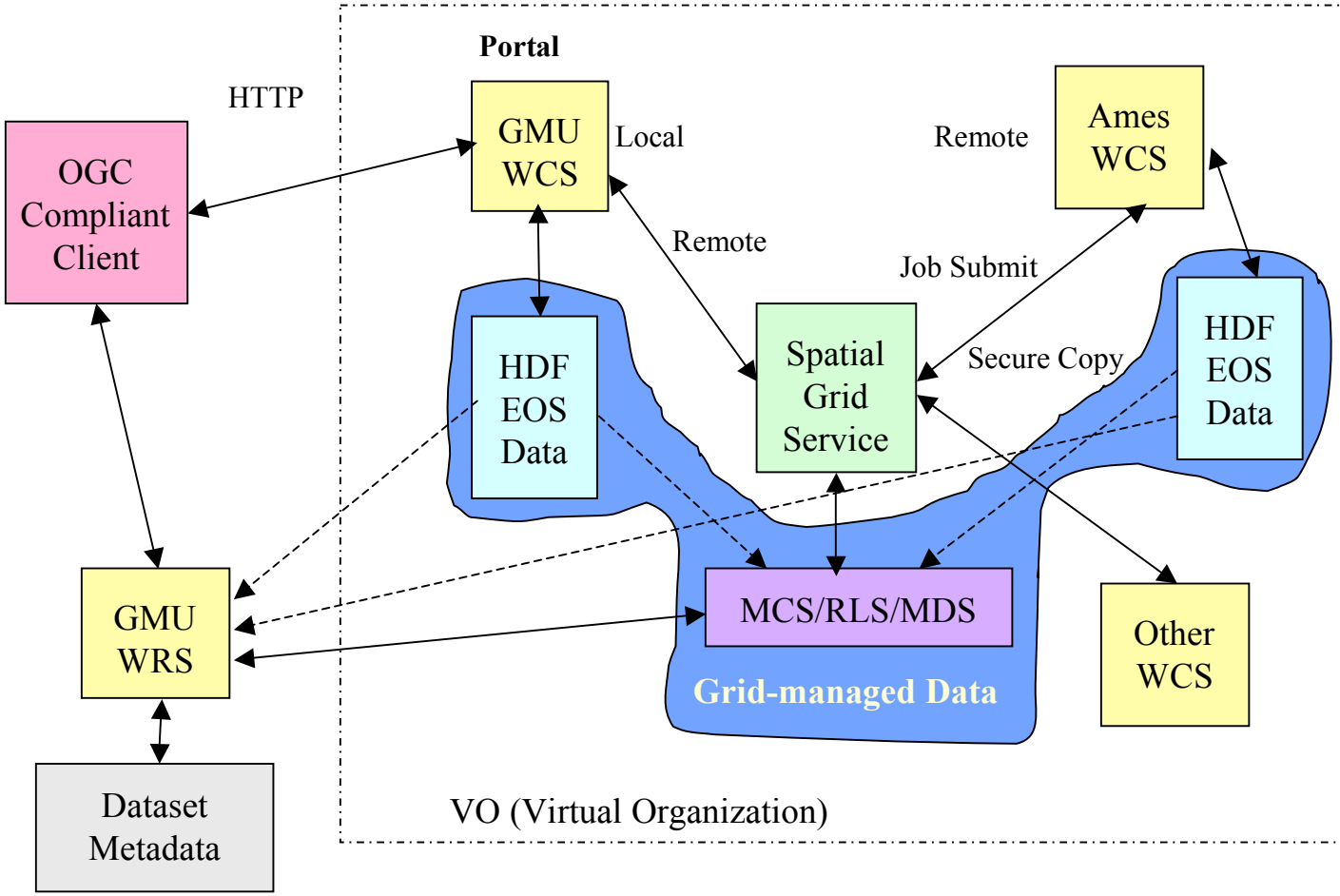
Testbed Environment Setup

- Installed gt3.0 on the Sun Solaris server and gt2.2 on the Linux server at GMU LAITS.
- Installed gt3.0 on Linux box and gt2.4 on the Irix server at NASA Ames.
- Issued IPG certificates to the above machines at GMU LAITS and NASA Ames, and test and debug their authentication to each other among those boxes.
- Installed MCS/RLS at GMU LAITS and NASA Ames.
- Installed MySQL server at GMU LAITS and NASA Ames.

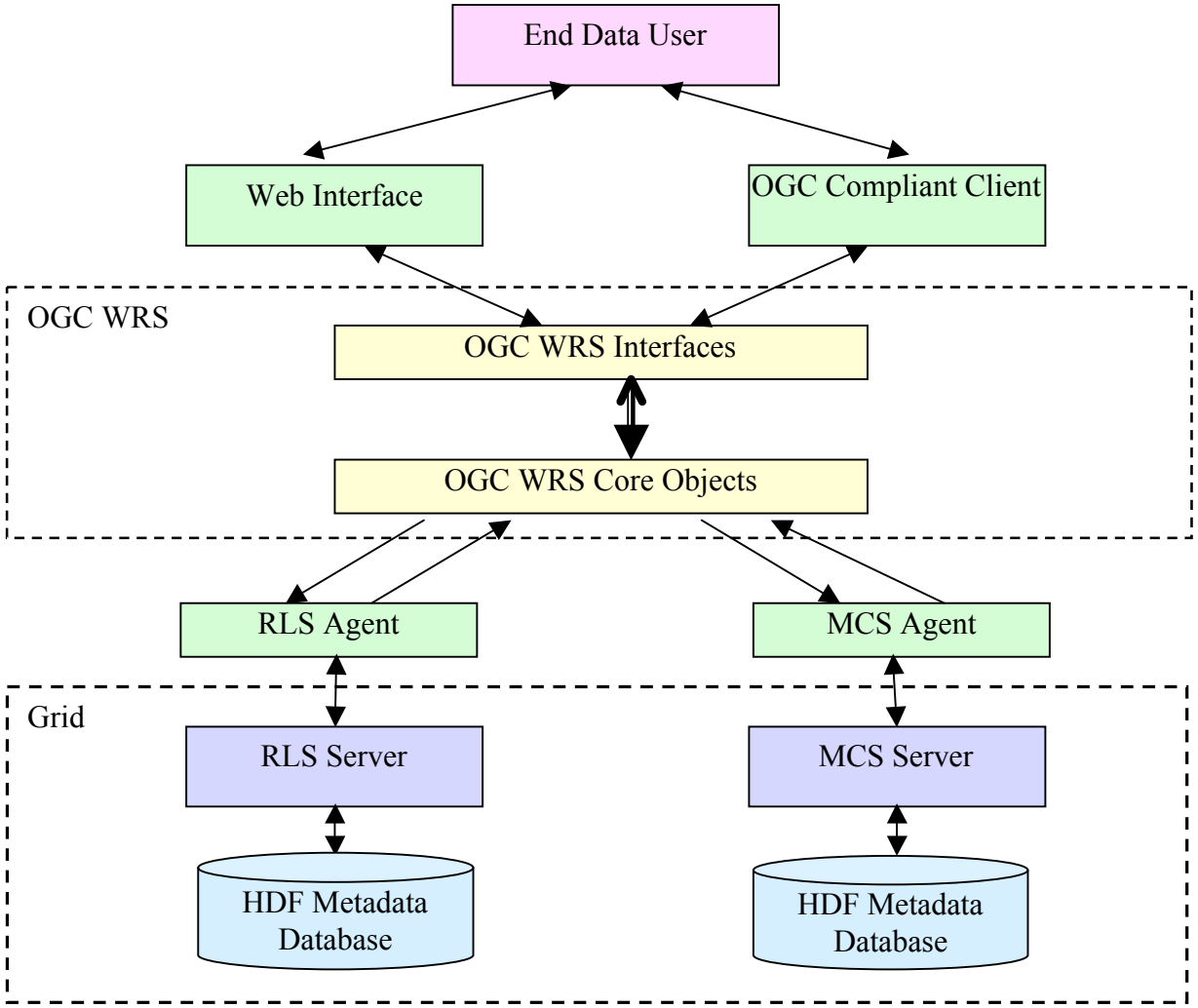
Grid Security and VO Setup



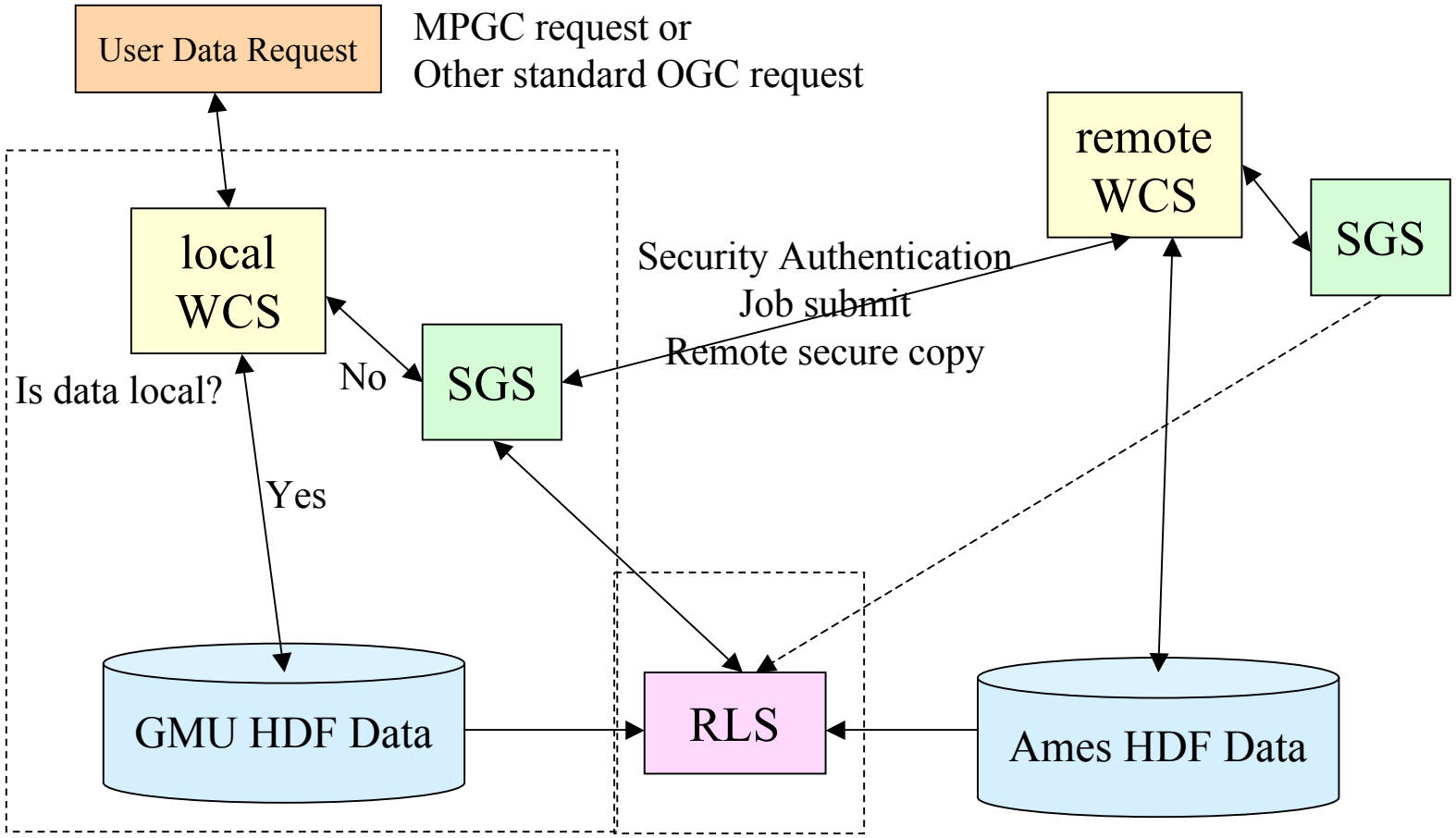
Integration Diagram for OGC & Grid



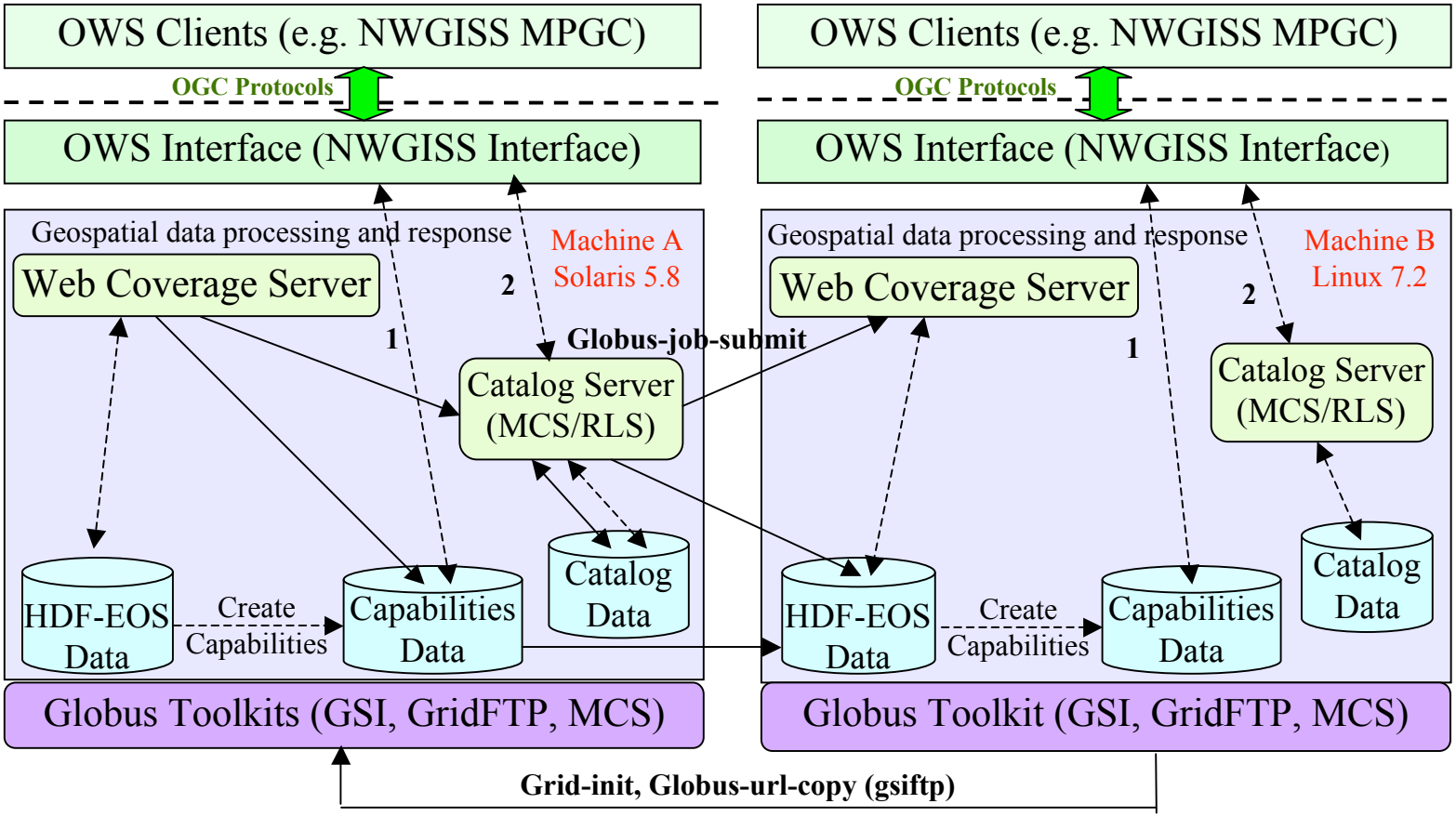
WRS and MCS/RLS with Metadata Database



WCS Adaptation and its relation with SGS



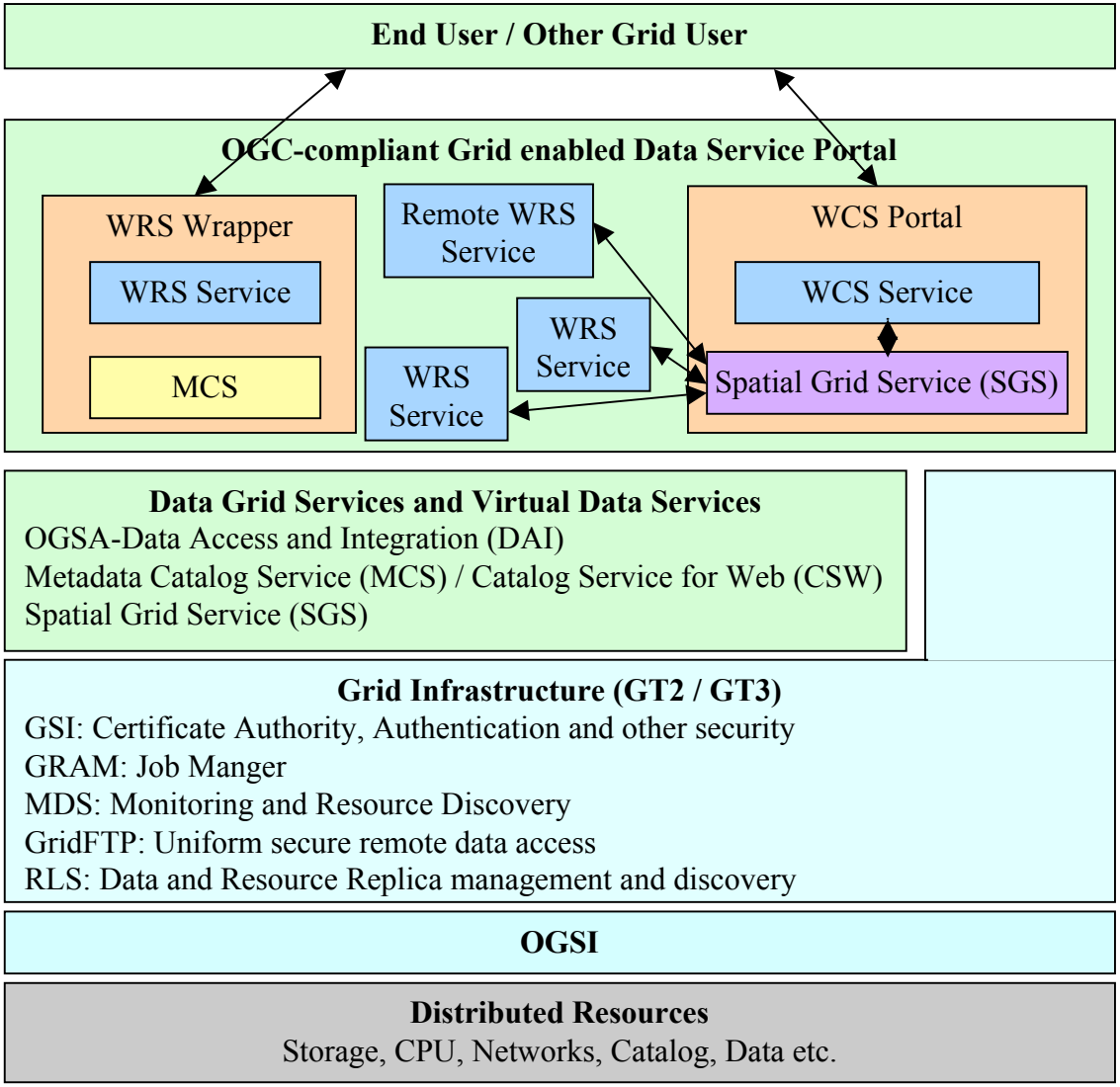
Data Flow Diagram for Integration



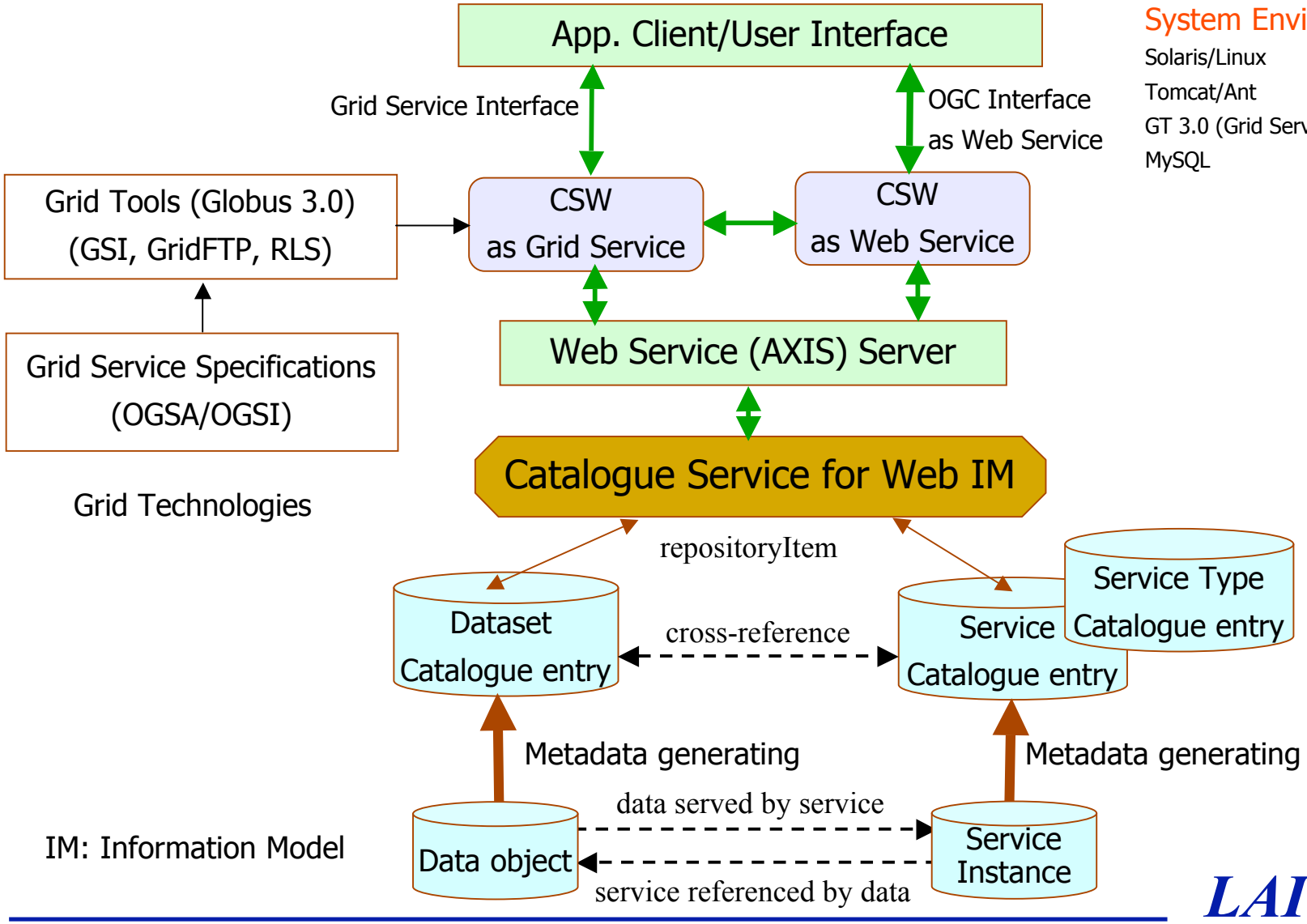
----- Broken lines show internal requests of NWGISS

—— Solid lines show requests related to Globus.

Components Architecture for Integration



Current implementation of OGC CSW



System Environment

Solaris/Linux

Tomcat/Ant

GT 3.0 (Grid Service)

MySQL

Future work

- We have basically finished the first two phases of the project;
- The major work currently is focusing on the enabling of the production of virtual geo-objects.

Work Plan for the Second 1/2

- May 2004 – Jul. 2004

  -- Install and master the latest stable Grid tools: gt3.2.

  -- Design the integration of current CSW with Grid (gt3.2).

  -- Advance the completed CSW to utilize the Grid technologies, also integrate our designed metadata catalogue into the Grid to contribute to the Grid technologies.

  -- Improve the current CSW as a Web Service to serve the Service Workflow and OGC standard client.

- Aug. 2004 – Oct. 2004

  -- Improve the completed Grid-enabled WCS to be based on Grid Services (gt3.2) for providing some new Spatial Data Services in gt3.2 environment, including secure data access, management, reliable file transfer, and so on. And integrate these services with the DAAC data pools.

Work Plan for Next Year 2/2

- Nov. 2004 – Jan. 2005

  -- Combine the CSW with the Replica Location Service (RLS) of Globus to provide a prototype of geospatial data replicas management for Earth Sciences, which will be completely transparent to all kinds of users.

  -- Partly add the semantic metadata of geospatial data and service into the CSW for query.

- Feb. 2005 – Apr. 2005

  -- Modify NWGISS WMS in the same manner as the WCS.

  -- Chain the CSW retrieval into the WCS/WMS operations in the Grid environment to serve the end user.

  -- Prepare to extend the CSW in Grid environment to support the registry and retrieval of the virtual geospatial data.